Contents

[**Q2-Aims and objectives:** 3](#_Toc37199858)

[**Brief description about the general aims of project:** 3](#_Toc37199859)

[**Objectives of above Aim:** 3](#_Toc37199860)

[**Q3-The Source Data:** 4](#_Toc37199861)

[**Source of data:** 4](#_Toc37199862)

[**Content of data:** 4](#_Toc37199863)

[**1.** **All Staff by equality chracteristics:** 4](#_Toc37199864)

[**2.** **Staff by mode of employement:** 4](#_Toc37199865)

[**3.** **All Staff by academic :** 4](#_Toc37199866)

[**Data Quality:** 4](#_Toc37199867)

[**Data Quality Table:** 5](#_Toc37199868)

[**Variable classification I:** 5](#_Toc37199869)

[**Q4-Research Question** 6](#_Toc37199870)

[**Al Staff Equality (Fig\_6) :** 7](#_Toc37199871)

[**Al Staff Academic (Fig\_4):** 7](#_Toc37199872)

[**Q5-Analysis and Findings:** 7](#_Toc37199873)

[**Answers:** 7](#_Toc37199874)

[**Ans 1** 7](#_Toc37199875)

[**Ans 2** 7](#_Toc37199876)

[**Ans 3** 8](#_Toc37199877)

[**Ans 4** 9](#_Toc37199878)

[**Ans 5** 9](#_Toc37199879)

[**Ans 6** 10](#_Toc37199880)

[**Ans 7** 11](#_Toc37199881)

[**Ans 8** 12](#_Toc37199882)

[**Ans 9** 13](#_Toc37199883)

[**Ans 10** 13](#_Toc37199884)

[**Q6-Project Description** 14](#_Toc37199885)

[**1-** **Project Planning:** 14](#_Toc37199886)

[**2-** **Data Aquiring:** 14](#_Toc37199887)

[**3-Preparation:** 14](#_Toc37199888)

[**4-Analyzing:** 14](#_Toc37199889)

[**5- Report of our findings:** 15](#_Toc37199890)

[**Q7–Reflection:** 15](#_Toc37199891)

[**Q8- References:** 16](#_Toc37199892)

**Executive summary**

We have been assigned the three related data sets of HESA, the Higher Education Statistics Agency, the experts in UK higher education data, and the designated data body for England.

The datasets provide detail information about staff of HESA. It contains details for their contract , country, description of zero-hour contract, part-time , full-time contract along with their responsibilities. We have shown the relationships between selected variables using appropriate visualizations with Pandas data frame in jupyter notebook. We explored all the three datasets but analyzed only two in detail. Also, we produced our conclusion according to 10 different questions.

# **Q2-Aims and objectives:**

## **Brief description about the general aims of project:**

* Learn by practicing how to use data sets.
* Learn by practicing how to select important features to find a problem and then its solution.
* Learn by practice some statistics in each data sets
* Learn how to use Jupiter notebook for analyzing and working with csv files in analysis by using pandas and NumPy libraries.
* Learn how to use visualization with matplotlib.

## **Objectives of above Aim:**

* Formulate a research question and investigate it and write a report of what we found.
* Import required libraries into ANACONDA(Jupyter notebook).
* To investigate the relationships between a selected independent variable and a selected dependent variable.
* Find the relationship between selected variables.
* Use SQL language and Pandas data frame.
* Use of a proper measure of correlation to show the relationships between our selected variables and appropriate visualizations using different libraries.

# **Q3-The Source Data:**

## **Source of data:**

All the data sets were taken from official website of <https://data.gov.uk/>.

## **Content of data:**

Basically, there are three related datasets we were assigned to picked and use for analysis. which are as follows.

### **All Staff by equality characteristics:**

This data set has information regarding Age Group, Age Category, Country of Higher Education Provider, Standard Occupation and having information of staff those who are on academic contract.

### **Staff by mode of employment:**

The data contains important information of Category of employment whether it is full time or part time , gender information and percentage.

### **All Staff by academic :**

This data has information of mode of employment whether part time/ full time , defines terms of employment, fixed term or permanent, identifies zero-hour contract , gender identification and hourly paid marker.

Hence The data sets contain all the detailed information related to higher education staff of United Kingdom according to their staff’s country, employer type, their activities and gender.

## **Data Quality:**

After getting data first we clean it by using python script (code) in Jupiter notebook.

First, we wrote code to checked up for missing and null values, after it we wrote code to check duplicated data and hence, we had not found any duplicated and missing data (values) in all three data sets. After it we use python code and identified some outliers in two data sets. 35 outliers were found in ‘Number’ feature of all staff equality data set and 25 outliers were found in ‘Number’ feature of all staff academic dataset. We have written code of it in Jupiter notebook as well.

After it we found some unformal data such as Sex column should not have any values ‘All’. Same as we found ‘All’ in zero-hour contract and Academic contract marker and category columns should not require total attributes. Same as we don’t require ‘All’ attribute in Mode of employment column and Terms of employment don’t require ‘All’ attributes.

All the python code has been placed in Jupiter notebook to show these dirty data. The dirty data was important to remove otherwise it can compromise our model.

## **Data Quality Table:**

|  |  |
| --- | --- |
| Term | Tracing and check |
| Validity | Data is perfectly valid. |
| Accuracy | There are some outliers for upper and lower bounds in Number columns of both data sets All staff equality and All staff academic respectively.  We use python code in notebook file to identify these outliers. |
| Completeness | There were no Null or missing values, so data set is complete. |
| Consistency | Data is perfectly consistence. |
| Uniformity | There are some unformal data such as ‘All’ attribute in ‘Sex’ column  Similarly, in ‘category’ column we have ‘total’ attribute and ‘mode of employment’ can’t be ‘All’. That’s why these were not uniform data. We determined with help of python code in Jupiter notebook. |

## **Variable classification I:**

On the behalf of viewing datasets, we are showing all three datasets where these are dependent or independent according to the nature of their attributes.

|  |  |
| --- | --- |
| All Staff Academic | Dependently |
| Academic contract marker | Independent |
| Terms of employment | Dependent |
| Zero hours contract | Dependent |
| Mode of employment | Independent |
| Sex | Independent |
| Hourly paid marker | Dependent |
| Number | Dependent |

|  |  |
| --- | --- |
| All Staff Equality | Dependently |
| Category Marker | Independent |
| Category | Independent |
| Country of HE Provider | Dependent |
| Academic contract marker | Dependent |
| Activity standard occupational classification | Dependent |
| Number | Dependent |

|  |  |
| --- | --- |
| Staff by Mode | Dependently |
| Category Marker | Independent |
| Category | Independent |
| Sex | Independent |
| Percentage | Dependent |
| Number | Dependent |

# **Q4-Research Question**

As stated in summary we will find ten different solutions in this project, so the list of our research questions is Below.

Q1. How many Males has Open-ended/permanent contract?

Q2. Find and Visualize the Total number of staff persons in each Mode of employment.

Q3. Count the Total number of staff persons in each Hourly paid marker using visualization as well

with bar graph?

Q4. Show the hourly paid salary for Females in Part-time contract?

Q5. How many staff members in each country with age category Asian and 26-35?

Q6. Find and visualize the Total count of staff in each category when country is England and Scotland.

Q7. Show the combine list of Total academic staff from Activity standard occupational classification

And England from Country of HE Provider. Also find the total rows when they both comes together.

Q8. Find and visualize total count of Activity standard occupational classification Staff.

Q9. Find the list of Staff of Non-academic contract when category is White.

Q10. Find and visualize total number of Male, Other and Females On a zero-hour contract.

Well as we are not doing only one question it’s a series of ten question so one by one to define everyone with dependent/independent would be a wrong attempt because in one question we may let it is dependent and may take independent in other. So here let me give the list of all variables who will be under investigate.

## **Al Staff Equality (Fig\_6) :**

These are the variables of this datasets who will be under investigation

|  |
| --- |
| Al Staf Academic (Fig 6) |
| Academic contract marker |
| Mode of employment |
| Terms of employment |
| Zero hours contract |
| Sex |
| Number |

## **Al Staff Academic (Fig\_4):**

The table below shows all the variables was under investigation of this data set.

|  |
| --- |
| Al Staf Equality (Fig 4) |
| Category Marker |
| Category |
| Country of HE Provider |
| Academic contract marker |
| Number |

Above are the table/list of variables those we had used/investigated in jupyter notebook to get answer of the ten question as wrote above here.

# **Q5-Analysis and Findings:**

## **Answers:**

We have used SQL queries, Numpy, Pandas and matplotlib for statical methods and analysis in jupyter notebook. Here we only mention the answers of above wrote ten questions.

**Ans 1**- We used the aggregated function to count the number of males in the open-

ended/permanent term contract and we found 601195 males using this type of contact.

**Ans 2**- we used SQL query and Statistical method sum to count total number of staff members

who are working on part time or full-Time .The list is as below.

Part-Time members are 575055.

Full-Time members are 1184735.

|  |
| --- |
|  |

After this question we found that there are too much Full-Time staff which is part of

HE then Part-Time staff.

#### **Visualization of Total Number of Staff in Each Mode:**

A screenshot of a cell phone

Description automatically generated

**Ans 3**- We used SQL query and statistic method to find total number of staff based upon

category of hourly paid ,total, other and salaried .

A screenshot of a cell phone

Description automatically generated

After finding the solution of this question we conclude that there is too much strength of staff

Are working as hourly paid and very less staff is Other category only are 465 in counting.

#### **Visualization of Total Number of staff In Hourly Paid Marker**

A screenshot of a social media post

Description automatically generated

**Ans 4**- We had combined the result by two columns and showing total number of females

working as Part-Time contract. And these are 36655 in counting.

A screenshot of a cell phone

Description automatically generated

**Ans 5**- we used SQL query to find the number of staff in each country with age category 26-35 and

Asian and found that

A screenshot of a cell phone

Description automatically generated

By this analysis we found that Northern Ireland has very less staff of category 26-35 and Asian.

#### **Visualization of Total Number of Staff Having Age Category 26-35 and Asian**

A screenshot of a cell phone

Description automatically generated

**Ans 6**- In this question we just use statistical method to find all the sum of all categories

results of total number of staff with all categories in England And Scotland.

A screenshot of a cell phone

Description automatically generated

It was too big picture so could not capture whole picture, but second column displays

count and Third contains category list. However, this analysis shows the difference

between numbers of Counting of all staff with all categories in England and Scotland.

#### **Visualization Total Staff in Each Category In England And Scotland**

A screenshot of a cell phone

Description automatically generated

**Ans 7**- In this We had to display total number of staff having Total academic staff and are of

England also we have showed there are 22 times happen in our data to see same

Combination. A screenshot of a cell phone

Description automatically generated

**Ans 8**- We are using statistical method to find total number of staff of Activity standard

occupational classification and found that staff of clerical and manual occupation

category is very less in counting which is 35. Moreover, picture shows the complete

number of difference between all activity category of staff.

A screenshot of a cell phone

Description automatically generated

#### **Visualization of Total Number of Staff In Activity Standard Occupation Classification**

A screenshot of a social media post

Description automatically generated

**Ans 9**- In this question we found out the relation when category is ‘White’

And all staff members of Non-Academic Staff and this combination repeated 50 times

in our data Set.

**Ans 10**- In this question we are exploring sex column to find total number of its attributes (Male, Other,

Female) with combination of On a Zero-hour contract.

After the analysis we found that with this query, counting of female is 7740 and male is 6155.

A picture containing clock

Description automatically generated

#### **Visualization of Sex Column with On a zero Hour Contract**

A screenshot of a cell phone

Description automatically generated

# **Q6-Project Description**

## **Project Planning:**

* Read the TMA project with each question requirement carefully.
* We made a schedule to complete each task in a specific time frame.
* For each question I had some notes and search in the web for some on things that I don’t understand.
* Divide all the tasks between all group members.

## **Data Acquiring:**

* We found all data from HESA website whose link is:

<https://data.gov.uk/dataset/452fa2dd-72e2-4de3-9e91-25be38dec27d/higher-education-staff-statistics-uk-2018-19.>

## **3-Preparation:**

* I downloaded all the three datasets from given link.
* Downloaded Anaconda
* Start coding on jupyter notebook

## **4-Analyzing:**

* Identifying variable in each data set.
* Checking the lookup table to find the relation between variables
* Checking the dependent and independent variables
* Answering different questions using python code to get the desired outcomes.

## **5- Report of our findings:**

Loaded the dataset. Remove unnecessary columns. Consider rows with combination of independent and dependent variables. Other Important things in this project

* Exploratory Data Analysis (EDA) with Visualization
* Data Pre-processing
* Data Analysis

# **Q7–Reflection:**

* Our experience with the project:

The experience regarding this project was quite amazing. I have a good experienced

and I’ve got learned many things. Also got a good experience in working as a group.

And the most Important thing scheduling tasks and committing to the schedule.

* Learned:

Got learned many things about python, Anaconda, jupyter notebook, SQL queries and analysis.

* what we went well:

Everything was good specially we did all task on time.

* what went wrong:

I have tried to use some dataframe functions and got many errors, then I have fixed them with a lot of efforts.

* and how can you benefit from this experience in future projects

I've learned things on Data Frame ,bar graphs and Histograms most importantly. So, it will benefit me afterwards.

# **Q8- References:**

|  |  |  |
| --- | --- | --- |
| Author | Type of search | Link |
| [**AnkitRai01**](https://auth.geeksforgeeks.org/user/AnkitRai01/articles) | Selecting rows in pandas dataframe on conditions.  I use stack overflow and other sites to search this as well. | <https://www.geeksforgeeks.org/selecting-rows-in-pandas-dataframe-based-on-conditions/> |
| **Brad Solomon** | Histograms | <https://realpython.com/python-histograms/> |
| Nadia Rivera | Cleaning Dirty Data with Pandas & Python | <https://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/> |
| Not seen on site | Removing Rows | <https://thispointer.com/pandas-skip-rows-while-reading-csv-file-to-a-dataframe-using-read_csv-in-python/> |

Well we did a lot of search behind this project and we were familiar from different sites hence could not remember all of them from where we got help. And most of help we took from stack overflow many time when struct with an error.